A Novel Semantic Similarity Based Technique for Computer Assisted Automatic Evaluation of Textual Answers

Udit Kr. Chakraborty[1], Samir Roy[2], and Sankhayan Choudhury[3]

[1] Department of Computer Science & Engineering
Sikkim Manipal Institute of Technology, Sikkim
[2] Department of Computer Science & Engineering
National Institute of Technical Teachers' Training & Research, Kolkata
[3] Department of Computer Science & Engineering
University of Calcutta, Kolkata
{udit.kc,samir.cst,sankhayan}@gmail.com

Abstract. We propose in this paper a unique approach for the automatic evaluation of free text answers. A question answering module has been developed for the evaluation of free text responses provided by the learner. The module is capable of automatically evaluating the free text response of the learner S_A to a given question Q and its model text based answer M_A on a scale [0, 1] with respect to the M_A. This approach takes into consideration not only the important key-words but also stop words and the positional expressions present in the learners' response. Here positional expression implies the pre-expression and post-expression appearing before and after a keyword in the learners' response. The results obtained on using this approach are promising enough for investing into future efforts.

Keywords: evaluation, learners' response, evaluation, keywords, pre-expression, post-expression.

1 Introduction

Evaluation is an important and critical part of the learning process. The evaluation of learners' response decides not only the amount of knowledge gathered by the learner but also contributes towards refinement of the learning process. The task requires the evaluator to have the required knowledge and also to be impartial, benevolent and intelligent. However, all of these qualities may not always be present in human evaluators, who are also prone to fatigue. Reasons similar to these and the requirement of performing the task of evaluation on a larger scale necessitates the implementation of auto-mated systems for evaluation of the learners' response. Such mechanized processes would not only be free from fatigue and partiality but also be able to evaluate across geographical distances if implemented in e-Learning systems whose importance, popularity and penetration is on the rise.

M.K. Kundu et al. (eds.), *Advanced Computing, Networking and Informatics - Volume 1*,
Smart Innovation, Systems and Technologies 27,
DOI: 10.1007/978-3-319-07353-8_46, © Springer International Publishing Switzerland 2014

However, the task of machine evaluation is easier said than done for reasons of complexity in natural languages and the lack of our abilities in understanding them. These reasons have given rise to the popularity of other types of assessment techniques namely multiple choice questions, order matching, fill in the blanks etc., which in spite of their own roles are not fully reliable for evaluation of fulfillment of learning outcomes. Whether the efficacy of questions requiring free text responses are more than the other types is debatable, but it is beyond contention that free text responses test the learners' ability to explain, deduce and logically derive apart from other parameters which are not brought out by the other types of question answering systems.

The problem with the evaluation of free text responses lie in the variation in answering and evaluation. Since the learners' response is presented in his unique style and words, the same answer can be written in different ways due to the richness in form and structure of natural languages. Another problem is in the score assigned to the answer since the score assigned can vary from one individual to another.

The computational challenge imposed by this task is immense, because, to determine the degree of correctness of the response the meaning of the sentence has to be extracted. The semantic similarity which is a means of finding the relation that exists between the meaning of the words and meaning of sentences also needs to be considered.

The work presented in this paper proposes an automated system that evaluates the free text responses of the learner. The approach is in deviation from the currently existing techniques in a few areas and considers not only important keywords but also the words before and after them. Unlike n-grams technique, the number of words before and after a keyword is not fixed and varies depending on the occurrence of the next keyword. The current work is limited to single sentence responses only.

2 Previous Work

Interest in question answering has shifted from factoid questions to descriptive questions [1], for reasons already discussed. A number of systems using different techniques have been developed for the evaluation of the free text response of learners namely: Intelligent Essay Assessor [2] developed by Landauer, Foltz and Laham based on Latent Semantic Analysis (LSA) [3] is used for the evaluation of learner essays. Apex a web based learning application developed by Dessus *et al.* [4] is also based on LSA technique and is used for the evaluation of learners' responses. Atenea developed by Perez *et al.* [5] is based on the Bi-Lingual Evaluation Understudy (BLUE) [6] Algorithm and is capable of evaluating the free text responses of learners in both English and Spanish irrespective of the language the learner wishes to answer. C-Rater [7] developed by Education Testing Service (ETS) uses Natural Language Processing (NLP) techniques for the evaluation of short responses provided by the learners. Auto-mark developed by Mitchell *et al.* [8] is a software system which uses NLP techniques and is capable of evaluating free text responses provided to descriptive questions.

However, a large scale acceptance of these systems, have not yet taken place and a complete replacement of the human evaluator is still a long distance away.

3 Proposed Methodology

The Answer Evaluation (AE) Module consists of two parts, one for the teacher and other for the learner. The role of the teacher would be to fix the model answer for a given question and fix the parameters of evaluation. This is similar to preparing a solution scheme of evaluation by the teacher which is referred to while actually evaluating the responses written by learners. The learners merely use the AE module to type in there textual responses to the questions presented.

The task chalked out for the AE module can be stated as:

Given a question Q, its model text based answer M_A and a learner response S_A, the AE module should be able to evaluate S_A on a scale of [0, 1] with respect to M_A.

- If the S_A is completely invalid or contradictory to M_A, then it is an incorrect response and a value 0 is returned.
- If the S_A is exactly same as the M_A or is a paraphrase of the M_A, or is a complete semantic match, then it is a correct response and a value 1 is returned.
- If the S_A is non-contradictory and is a partial semantic match for M_A, then the response is partially correct and a value greater than 0 and less than 1 is returned depending upon the match.

This work is built upon the understanding that, an answer to a question is a collection of keywords and their associated pre and post expressions which augment sense to the keywords in the context of the question and also establishes links between them. Unlike the nugget approach, which considers only keywords as the building blocks, the current approach considers the preceding and following sets of words as well. The choice of pre and post-expression is not based on the popular n-gram technique but the occurrence of the next keyword. It is also worth mention that unlike other natural language processing approaches, we do not remove the stop words from a response as we consider these to be important information carriers. Fig.1. presents the idea of how the answers are perceived by the model.

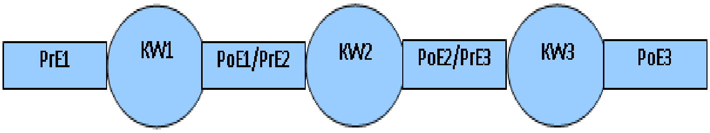

Fig. 1. Schematic diagram of the models perception of an answer

Since the system deals with natural language answers, we do not decide or attach any weight to the order in which the key-words appear while evaluating the responses. Also, it may be possible that a particular part of the response acts as post-expression for a keyword and pre-expression for the next keyword, in which case it will be considered twice depending upon the solution scheme presented by the teacher.

Each pre-expression and post-expression is again broken up into four parts, namely logic, certainty, count and part-of, which are the expected types of senses that these expressions attach to the keywords. There is however, no fixed order in which the words belonging to any of these categories would appear and it is also possible that they do not appear at all. Lists of words have been pre-pared to be belonging to each of these four categories and they are as shown in Table 1, 2, 3 and 4 respectively.

Table 1. Logic Expressions and their associated logic

S.No.	Logic Expressions	Logic
1	and	Conjunction
2	or	Disjunction
3	either-or	Exclusive Disjunction
4	only if	Implication
5	if and only if	Equivalence
6	just in case	Bi-conditional
7	not both	Alternative Denial
8	neither – nor	Joint Denial
9	not it is false that it is not the case that	Negation
10	is	Equality

Table 2. Certainty Expressions and their classes

S.No.	Certainty Expressions	Certainty Classes
1	usually, likely, unlikely	Class A
2	certainly, most certainly, definitely	Class B
3	most probably, most likely	Class C
4	probably	Class D

Table 3. Count Expressions and their meanings

S. No.	Count Expressions	Count
1	a, an, the, it, that	SINGULARITY
2	couple	DUALITY
3	those	MULTIPLE

Table 4. Part-of Expressions

S. No.	Part of Expressions
1	belong
2	into
3	for
4	like

3.1 Preprocessing Tasks

Prior to the evaluation process, the teacher has to perform the following preprocessing tasks to prepare the system to be able to evaluate the learners' response. This comprises of typing in the model response, identifying the model phrase from the complete response, identifying the keywords, the post and pre expressions for each key word and the categorization of the words in post and pre expressions into their rightful sense conveying types.

The model answer is the answer prepared by the human evaluator and presents the benchmark against which the learners' response would be evaluated. This answer consists of a central part, which we call the model phrase, M_P, and represents the core of the answer.

To ensure that the reproduction of your illustrations is of a reasonable quality, we advise against the use of shading. The contrast should be as pronounced as possible.

If screenshots are necessary, please make sure that you are happy with the print quality before you send the files.

The steps are listed below in the order of their occurrence:

Step 1: The model answer M_A is created.
Step 2: A model phrase M_P is identified within the model answer M_A.
Step 3: Keywords are identified and listed.
Step 4: All KW are marked with their associated part of speech.
Step 5: For each KW:
    Step 5.1: Synonyms having same POS usage are listed.
Step 6: Weights are associated to each KW depending on importance and relevance. Sum of all weights to be equal to 1.
Step 7: For every KW the pre-expression and the post-expression are extracted and words/phrases put in their respective sense brackets, i.e., logic, certainty, count or part-of.

3.2 Steps for the Evaluation of the Learner Response

Once the pre-processing is done, the system is ready to read in the learners' response and evaluate it. The aim is to evaluate the response and return a score in the range of 0 to 1. The algorithm for performing the same is as follows:

Algorithm Eval_Response:
Evaluates the learners' text based response.

Variables:
S_A (Learner's Response), M_A(Model Answer), M_P (Model Phrase), KW (Keyword), PrE(Pre-expression), PoE (Post-expression), KW_S(Score from a keyword), KW_Weight (Weight of a keyword), PrE_S(Score of a pre-expression), PoE_S(Score of a post-expression), Marks(Total marks).

```
Step 1: Set Marks = 0
Step 2: String Compare ($S_A$,$M_A$)
Step 3: If ($S_A$=$M_A$), Marks = 1
Step 4: ElseIf (Search ($S_A$, $M_P$) = Success), Marks = 0.85 //search the learners'response for the model phrase.
Step 5: Else For every KW Search (KW,$S_A$ ) //search for the keyword in $S_A$
    Step 5.1: If KW found: Evaluate (PrE, PoE) //Evaluate both pre and post-expressions
    KW_S = PrE_S*KW_Weight*PoE_S
    //Score of keyword is calculated
    Step 5.2: Else Search every synonym of current keyword in $S_A$.If synonym match found.
GOTO Step 5.1, Else KW_S = 0
        Step 5.3: Marks = Marks + KW_S
Step6: Stop
```

While evaluating the pre-expression (and the post-expression), we search whether the words listed in the solution scheme appear in the pre-expression. If they do, then they are put in the sense bracket that they are expected to belong. Otherwise, we check whether a valid substitution of the word appears in the pre-expression of the learners' response. The valid substitutions of logic, certainty, count or part-of word is maintained in a list along with their weights. For example, if the pre-expression in the solution scheme shows that the count expression is 'a' and the learners' response contains no 'a' in the pre-expression but a 'the', then the count expression for that particular keyword is taken as 'the' in the learners' response and is given a score of 0.5 instead of 1 (had it been 'a'), because 'a' signifies 'one in many', while the word 'the' signifies 'the only one'.

Each sub-field, namely logic, certainty, count and part-of, in the solution scheme may not be filled. Under such circumstances, we do not search for those expressions in the learners' response and such non-active fields do not contribute to the score. The exception to this rule is the logic sub-field, which contributes 1 to the score if, either the logic is inactive or the sub-field is active and the word is found in the learners' response. If the logic sub-field is active in the scheme, but the word is not present in the learners' response, the contribution becomes 0. If the field is active but the word is substituted, then the score changes appropriately.

Finally, every pre-expression and post expression score (PrE_S/PoE_S) is evaluated according to the expression given in Eq.1

$$\text{PrE\_S} = \frac{Logic*(Certainty+Count+Part\_of}{No.of\ active\ sub\_fields} \qquad (1)$$

4 Experiments and Results

The methodology discussed in the previous sections was employed to test the correctness in comparison to a human evaluator. While performing the tests, we considered single sentences responses only. The human evaluators were kept unaware of the method to be employed by the automated system; however, since the automatic evaluation would return fractional values between 0 and 1, human evaluators were asked to score up to 2 decimal places. Two such tests and their details are presented here along with some findings on the results.

4.1 Set 1

Question: What is an annotated parse tree?
Model Answer: A parse tree showing the attribute value at each node is called an annotated parse tree.
Model Phrase: A parse tree showing attribute value at each node.

The question was presented to 39 learners', and the responses evaluated by the automated system and also by human evaluators, based on the model response specified and shown in Table 5. The correlation co-efficient between the two evaluators was calculated and found to be equal to 0.6324, with 30% cases having difference not more than 10%.

4.2 Set 2

Question: What is the advantage of representing data in AVL search tree than to represent data in binary search tree?
Model Answer: In AVL search tree, the time required for performing operations like searching or traversing is short. e.g. worst case complexity for searching in BST (O(n)) worst case complexity for searching in AVL search tree (log(n)).
Model Phrase: In AVL search tree, the time required for performing operations like searching or traversing is short.

This question was asked to a group of 50 learners' and the responses similarly evaluated with model response as in Table 6. The results found returned a correlation co-efficient of 0.6919, with 68% instances of not more than 10% difference in the marks allotted by the system and the human evaluator.

It is observed that the performance of the system tends to improve on increasing the volume of the response. However, the in-crease in the volume of the response in this case also meant an increase in the number of keywords. As a matter of fact, both tests were conducted on keyword heavy samples. Whether the performance would change on having heavier pre and post expressions is yet to be explored.

Table 5. Scheme for the evaluation of Set 1

Pre-Expression				KW	Wt.	Post-Expression			
L	C	O	P			L	C	O	P
-	-	A	-	parse	0.2	-	-	-	-
-	-	-	-	tree	0.05	-	-	-	-
-	-	-	-	showing	0.2	-	-	the	-
-	-	-	-	attribute	0.3	-	-	-	-
-	-	-	-	value	0.05	-	-	-	-
-	-	each	-	node	0.2	-	-	-	-

Legends: L: Logic; C: Certainty; O: Count; P: Part of

Table 6. Scheme for the evaluation of Set 2

Pre-Expression				KW	Wt. of	Post-Expression			
L	C	O	P		KW	L	C	O	P
-	-	the	-	time	0.1	-	-	-	-
-	-	-	-	Required	0.1	-	-	-	-
-	-	-	for	Performing	0.05	-	-	-	-
-	-	-	-	Operation	0.05	-	-	-	like
-	-	-	-	Searching	0.1	or	-	-	-
-	-	-	-	Traversing	0.1	-	-	-	-
is	-	-	-	short	0.2	-	-	-	-
-	-	-	for	e.g.	0.01	-	-	-	-
-	-	-	-	worst	0.01	-	-	-	-
-	-	-	-	case	0.01	-	-	-	-
-	-	-	-	Complexity	0.015	-	-	-	-
-	-	-	for	Searching	0.01	-	-	-	-
-	-	-	in	BST	0.03	-	-	-	-
-	-	-	-	(O(n))	0.07	-	-	-	-
-	-	-	-	worst	0.01	-	-	-	-
-	-	-	-	case	0.01	-	-	-	-
-	-	-	-	Complexity	0.015	-	-	-	-
-	-	-	for	Searching	0.01	-	-	-	-
-	-	-	in	AVL	0.02	-	-	-	-
-	-	-	-	search	0.005	-	-	-	-
-	-	-	-	tree	0.005	-	-	-	-
-	-	-	-	(log(n))	0.7	-	-	-	-

5 Conclusion

The work is aimed at developing a novel system that is capable of evaluating the free text responses of learners'. Unlike the widely followed bag-of-words approach, the work mentioned here takes positional expressions, keyword and even stop words into consideration during evaluation. The proposed method generates a fuzzy score taking into consideration all mentioned criteria. The score generated by the system does not deviate too much from the human evaluator and further tests may produce still better results.

References

1. Lin, J., Fushman, D.D.: Automatically Evaluating Answers to Definition Questions. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pp. 931–938 (2005)

2. Foltz, P.W., Laham, D., Landauer, T.K.: The Intelligent Essay Assessor: Applications to Educational Technology. Interactive Multimedia Education Journal of Computer Enhanced Learning 1(2) (1991)
3. Landauer, T.K., Foltz, P.W., Laham, D.: An Introduction to Latent Semantic Analysis. Discourse Processes 25(2&3), 259–284 (1998)
4. Dessus, P., Lemaire, B., Vernier, A.: Free-text Assessment in a Virtual Campus. In: Proceedings of the 3rd International Conference on Human-Learning Systems, pp. 2–14 (2000)
5. Perez, D., Alfonseca, E.: Adapting the Automatic Assessment of Free-Text Answers to the Learners. In: Proceedings of the 9th International Computer-Assisted Assessment (CAA) Conference (2005)
6. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318 (2002)
7. Leacock, C., Chodorow, M.: C-rater: Automatic Content Scoring for Short Constructed Responses. In: Proceedings of the 22nd International FlAIRS Conference, pp. 290–295 (2009)
8. Mitchell, T., Russell, T., Broomhead, P., Aldridge, N.: Towards Robust Computerized Marking of Free-Text Responses. In: Proceedings of 6th International Computer Aided Assessment Conference, Loughborough (2002)